Tune Down the Misinformation, Please: Generating Corrective Messages for COVID-19 Misinformation

Dylan Meyer
Fairfield University
Fairfield, CT, USA, 06824
dylan.meyer@student.fairfield.edu

Jie Tao
Fairfield University
Fairfield, CT, USA, 06824
jtao@fairfield.edu

Alison E. Kris
Fairfield University
Fairfield, CT, USA, 06824
akris@fairfield.edu

Abstract

The ongoing COVID-19 pandemic drastically changed our lives in multiple aspects, one of which is the reliance on social media during quarantine, both for social interaction and information-seeking purposes. However, the wide dissemination of misinformation on social media has impacted public health negatively. Previous studies on COVID-19 misinformation mainly focused on exploration of impacts and explanation of motivations, with few exceptions. In this study, we propose an analytical pipeline that generates corrective messages toward COVID-19 misinformation in a semi-automatic fashion, and then evaluate it against a large amount of data. Both the automated and manual evaluation results suggest the efficiency of the proposed pipeline, which can be used in combination with human intelligence by individuals and public health organizations in fighting COVID-19 misinformation.

1. Introduction

The ongoing pandemic drastically impacted the lives of millions and society at large. During this time, many rely on social media, not only for social interaction purposes, but also as a primary information source. Compounding this, since no prior information regarding COVID-19 existed, the general population is more likely to regard social media as their main information source [1]. Although much information on social media may be trustworthy, the reliance of social media use in our daily lives has also contributed to the wide dissemination of misinformation related to COVID-19, due to the lack of rigorous critiques as well as ubiquitous radical ideas and misconceptions [2], which is also termed as information disorder in the literature [3]. Misinformation can be a matter of life and death amid a public health crisis like COVID-19, compared to other types of misinformation. A recent survey suggested that 48% of Americans had encountered misinformation regarding COVID-19 [4]. Additionally, another study by Loomba et al. [5] found that exposure to COVID misinformation lowered the number of participants who would 'definitely' take the vaccine by 6.2%. Consequently, researchers and practitioners have studied and quantified the severity of COVID-19 misinformation. Singh et al. [6] conducted an exploratory study toward the metadata of COVID-19 misinformation on Twitter, including volume, location and key terms/topics from 18 million collected tweets discussing COVID-19. Kouzy et al. [7] collected a sample of 673 tweets for an exploratory analysis on COVID-19 misinformation and metadata (e.g., account information, number of followers/likes/retweets). Beyond meta-data analysis, researchers have analyzed social media content for misinformation detection and classification purposes. For instance, Hossain et al. [8] developed a dataset named CovidLies, which associates tweets to COVID-19 misconceptions, relying on the cosine similarity of word embeddings using the BERTSCORE [9] model without fine-tuning. However, the findings from previous studies are limited in several aspects. Firstly, most extant studies on COVID-19 misinformation focus on exploration (e.g., using the metadata and correlation analysis [7]) or explanatory purposes (e.g., hypothesis testing on impacts). Secondly, few studies (e.g., [1]) focus on the detection of COVID-19 misinformation. However, the detection performances are limited due to limited-sized datasets used in these studies. Thirdly, other studies utilized topic modeling techniques, which are unsupervised and less capable of capturing linguistic nuances involved in COVID-19 misinformation. However, prior studies have highlighted directions for future research, which include:

- Developing intelligence tools using machine learning and Natural Language Processing (NLP) for detecting and intervening against COVID-19 misinformation [10];
- Collecting data from multiple sources to improve the generalizability of the detection models [11];
- Providing suggestions for combating COVID-19 misinformation, if definition responses are not directly available [12];
- Leveraging prior domain knowledge, e.g., named entities (e.g., person, organization, location), which

URI: https://hdl.handle.net/10125/79369
978-0-9981331-5-7

are associated with 70% of COVID-19 misinformation [11].

In this study, we propose a Generative Adversarial Network (GAN) based pipeline to enhance detection and facilitate intervention toward COVID-19 misinformation. GAN is a generative model consisting of a generator and a discriminator, where the former generates data based on the distribution of the training data, and the latter distinguishes generated data from the original data [13]. However, as the generators are typically initialized with random distributions, much of the generated data is of lower quality and is difficult to be used for downstream analysis. Therefore, it is beneficial to employ prior domain knowledge in training GANs.

This study makes multi-fold research and technical contributions. Firstly, this is the first study to extend the analysis of COVID-19 misinformation in social media to the intervention level. Specifically, this study proposes a semi-automatic approach to generate corrective messages toward COVID-19 misinformation. Secondly, the generated texts from the proposed Masked Language Model (MLM-GAN) model can be used for data augmentation to further enhance the performance of detection models. Thirdly, we extend the existing GAN models by incorporating prior knowledge (e.g., Named Entities (NEs) and key terms), to improve the effectiveness and quality of the generated texts. Lastly, we include humans in the analytical loop by adding a manual review phase downstream from the misinformation intervention module. We collect a large amount of COVID-19 misinformation data from multiple sources to improve the generalizability of the proposed analytical pipeline. Both the automatic and manual evaluation results show clear evidence of the efficacy of the proposed pipeline, not only concerning detection but also intervention towards COVID-19 misinformation. We also report several observations, which identify future research directions regarding studies of misinformation handling. The analytical pipeline and the findings in this study can improve the collaboration between human users in public health organizations and intelligent systems, in combating misinformation on social media.

2. Prior Studies

2.1. Misinformation in COVID-19

Even before the COVID-19 global pandemic, social media had become the most popular venue for disseminating misinformation, which is partially due to the lack of critical thinking and the amplification of radical ideas in virtual communities [2]. Previous studies have also focused on exploring the impact and/or explaining the motivation of COVID-19 misinformation. For instance, Kouzy et al. [7] designed a correlation-based analysis on the metadata of COVID-19 misinformation tweets, and discovered that tweets from verified accounts and healthcare organizations had the lowest rates of misinformation. Additionally, Krause et al. [12] analyzed the multi-layered risks of COVID-19 misinformation when disseminated to the general public. Furthermore, Barua et al. [14] observed that although social media platforms embed fact-checking mechanisms to detect and possibly ban misinformation, it is still difficult to stop its spread.

The literature has suggested that intelligent systems need to be developed to support public health organizations (e.g., World Health Organization) to compose corrective messages in combat of COVID-19 misinformation [1], [2], [10], [11] in social media. For example, Tasnim et al. [10] suggested that machine learning and NLP tools should be leveraged for detecting and correcting COVID-19 misinformation. Choudrie et al. [1] built a detection model on 143 labeled data points, with the best model in this study yielding an 86.7% accuracy. Another school of study focused on topic modeling concerning COVID-19 misinformation. For instance, Hossain et al. [8] aligned COVID-19 related tweets with well-established misconceptions via a similarity-based method, and transformer-based models. Studies using similar methods can be found in [15], [16]. Despite the findings from these studies, they can be extended along following directions: 1) larger datasets can enhance the generalizability of the results and findings, 2) more advanced models that are capable of capturing the linguistic intricacies within the misinformation, and 3) Utilization of domain-specific knowledge for detection of COVID-19 misinformation.

2.2. Transformers and Generative Models

Transformer-based models are state-of-the-art developments in the field of NLP. Compared to traditional NLP methods (e.g., word2vec [17]), these models are more capable of capturing linguistic intricacies, since they are pre-trained with vast amounts of domain-independent, general purpose textual data for pseudo classification tasks such as token classification or language understanding. These models include BERT (Bidirectional Encoder Representations from Transformers) [18] and XLNet [19], which utilize the encoder component of the multi-layered transformer-based models. In order to achieve superior performance on texts from a certain domain, these models need to be *fine-tuned* with supervised domain adaptation, which entails the "re-training" of the models using domain-specific texts.

On the other hand, GPT-2 [20] is a multi-layered transformer decoder network, which is initially pre-trained using unlabeled data from 8 million webpages. Compared to transformer encoder models, which are typically used for text classification purposes, GPT-2

models are used to generate texts for specific purposes. For instance, Niewinski et al. [21] proposed a generative enhanced model to generate malicious claims from online articles for adversarial attacks on fake claim classifiers. Another type of generative models are MLMs. Compared to the GPT-2 generative models, which generate the remainder of the sentence based on seed words, MLMs generate sentences by substituting masked tokens with the most similar tokens from the language space. For example, Wu et al. [22] developed a MLM model to transfer the sentiments within sentences. However, when solely using the GPT-2 or MLM models for generation purposes it is difficult to control the quality of the generated content.

One of the most popular ways of addressing the aforementioned limitations is the GAN model [13]. Each GAN model consists of a generator, which captures the distribution in the data, and a discriminator, which estimates whether a data point is from the generator or original text. The training purpose of GANs is to maximize the probabilities of the discriminators to make mistakes, so that the generated texts from the generators are similar to the original training data. GANs have recently been applied to textual data. For example, Zhang et al. [23] proposed a GAN model consisting of a Long Short-Term Memory (LSTM) network as the generator, and a Convolutional Neural Network (CNN) as the discriminator, to generate adversarial examples to improve the robustness of text classifiers. One issue with this approach is that the LSTM based generators and the CNN-based discriminations are not sufficient to capture the contextual information in the contexts. Thus, researchers have proposed using transformer-based models as generators/discriminators. For instance, Irissappane et al. [24] designed a GAN model to generate spam reviews for augmenting their labeled training data. Utilizing transformer-based models (e.g., GPT-2) in GANs is a relatively new method and can be enhanced via: 1) using domain knowledge, rather than random distribution, to improve the efficacy; 2) fine-tuning to enhance the domain relatedness of the generated texts.

3. The Analytical Pipeline

The analytical pipeline proposed in this study, including preprocessing, modeling and post processing steps, contains several design novelties. First, We extend the literature by exploring the intervention mechanics with regard to COVID-19 misinformation. In particular, the misinformation intervention module provides either corrective suggestions (which require further manual editing and filtering), responses to misinformation (which requires minimal direct human intervention), or complete correction (no direct human intervention needed). Second, we employ state-of-the-art transformer-based models, specifically BERT and GPT-2, for the purpose of COVID-19 misinformation detection and intervention. Our proposed detection method is supervised, which streamlines the evaluation processes, and reduces the human intervention required in the evaluation of *candidate corrective messages*. Third, we design a teacher-forcing method improves the performances of both the detection enhancement and misinformation intervention modules. In general, a teacher-forcing algorithm is typically used to train recurrent models with input data and previous state [25]. In this study, we design the teacher-forcing method using prior human knowledge along with the training data in the training process of the GAN models, specifically:

- For the detection enhancement module, the teacher-forcing mechanism targets specific categories of NEs and other key terms and phrases most related to COVID-19. This mechanism allows for the generation of augmented training data (i.e., misinformation) to enhance the classifier for the purpose of detection.
- For the misinformation intervention module, the same entities and key terms are extracted from true information and added as additional input signals. In this manner, the model is trained to generate less misinformative texts, while maintain semantic relatedness to the original texts. Thus, the misinformation intervention module can generate less misinformative substitutes that are related to the themes of the original texts. Additionally, we design a rating schema (see Figure 3) to evaluate how thoroughly the generated texts respond to the original misinformation. The proposed schema can be used in other related studies to assess their impact in addressing COVID-19 misinformation in social media, or misinformation in general.

The proposed analytical pipeline contains three main modules, namely misinformation detection, detection enhancement, and misinformation intervention. The proposed analytical pipeline is depicted in Figure 1. In the misinformation detection module, we design the misinformation Detection Model based on a fine-tuned BERT for sequence classification model serves as the initial classifier for COVID-19 misinformation detection. This model is then enhanced in downstream processes with additional training data sourced from the Detection Enhancement module. The core of the detection enhancement module is a MLM-based GAN generates additional misinformative examples to enhance the performance of the Misinformation Detection Model. This serves a similar role to data augmentation tasks in computer vision analysis, where variations of images (e.g., via rotation, zooming in/out) are used as additional examples to train a more robust and highly performing

model. In the misinformation intervention module, a GPT2-based GAN generates whole texts in response to COVID-19 misinformation, bolstering human correction and validation efforts. These messages range from corrective suggestions (i.e., candidate messages containing true information but require human validation) to responses (i.e., direct and fully validated counterpoints to the original text), reducing the overall time complexity of manually addressing misinformation pertaining to the global pandemic.

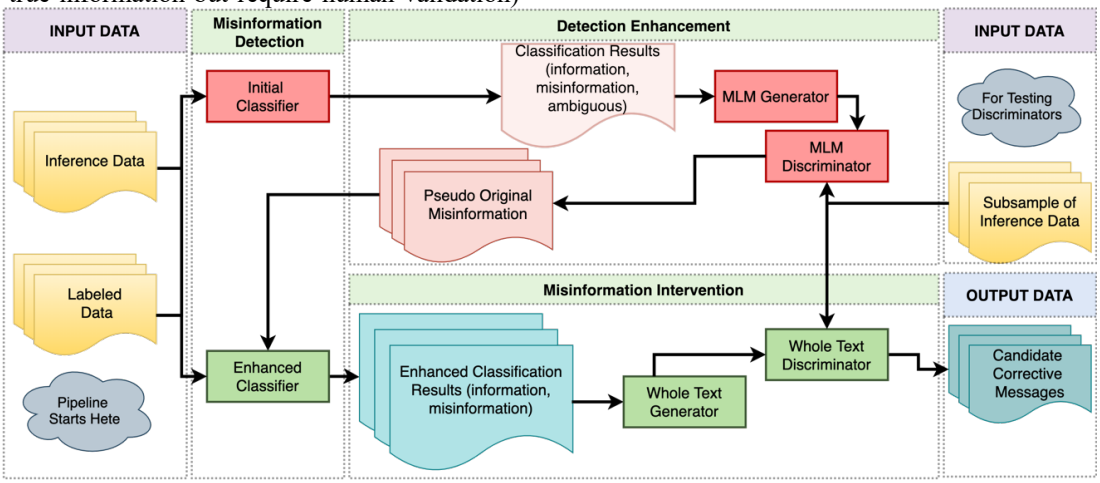

Figure 1. The Proposed Analytical Pipeline

3.1. Misinformation Detection

For both the initial and enhanced classifiers, we first fine-tune a pre-trained transformer model using labeled COVID-19 misinformation data. The fine-tuning step updates all model parameters to accommodate the linguistic intricacies from the domain, rather than the generic texts used to pre-train them. Given labels $y = \{0,1\}$, where 0 indicates misinformation and 1 indicates true information, the objective of updating the model parameters is to minimize the loss in identifying misinformation. Let X denote a COVID-19 related social media post (misinformation or true information) with m tokens. Each token is then embedded by the model as a n-dimensional vector. The transformer model embeds the texts through all the layers, and we then select the row d_{CLS}, which corresponds to the [CLS] special token added to the post, from the embedding matrix $D \in R^{m \times n}$, to calculate the sentence embedding s via eq. (1), where $W \in R^{m \times |y|}$ and $b \in R^{|y|}$:

$$s = d \cdot W + b \quad (1)$$

With the sentence embedding learned, misinformation detection is designed as the downstream task. To use transformer models as classifiers, we add a SoftMax layer as the output layer, which yields a probability distribution $P(y|s)$ as the SoftMax function over s. For example, if X is misinformation, $P(y = 0|s)$ is calculated via eq. (2):

$$P(y = 0|s) = softmax(s) = \frac{e^{s[0]}}{\sum_y e^{s[y]}} \quad (2)$$

In the output layer, we use the negative log-likelihood loss as the objective function (see eq. (3)):

$$J(X) = -log(P(y|s)) \quad (3)$$

As mentioned previously, the purpose of the fine-tuning process is to minimize $J(X)$. Additionally, we use the sentence embeddings $S, s \in S$, in traditional Machine Learning classification models (e.g., eXtreme Gradient Boosting, Random Forest) to perform the same classification task. The classifier with the best classification performance is then selected, and used to classify the *inference data* (i.e., the unlabeled data), to distinguish misinformation from true information. It is worth noting that in addition to information and misinformation, we select the data with relatively low classification confidence from the initial classifier (i.e., classification probabilities between 0.4 and 0.6), termed as *ambiguous data*, for the purpose of evaluating whether the misinformation detection is enhanced using the detection enhancement module discussed in Section 3.2. Furthermore, the output from the enhanced classifier is then used as input for the generator in the misinformation intervention module. We hypothesize that after employing the enhanced classifier, more instances in the inference data are classified with higher probabilities toward true information and misinformation (i.e., less ambiguous), which is evaluated in Section 4.2.

3.2. Detection Enhancement

In the detection enhancement module, we design and implement an MLM-based GAN model. The purpose of the MLM-GAN model is to generate variants of

texts that are semantically similar to the original unlabeled texts in the inference data, so that the detection performance on the ambiguous data is improved in the downstream enhanced classifier [22]. The MLM generator is trained on a random sample of the inference data that are classified as misinformation by the initial classifier and combined with the sample of labeled misinformation of equal size. This allows the model to generate variants of misinformation from original texts. We target misinformation generation in this module, as the identification of misinformation is more vital than true information due to the potentially inflammatory and dangerous notions expressed in widespread misinformation, which can potentially threaten public health. In addition, rather than masking a random subset of tokens from each text, we mask specific types of NEs and the most important key terms (i.e., based on their Term Frequency-Inverse Document Frequency (TF-IDF, a statistical measure that evaluates how relevant a term is to a social media post in respective datasets, e.g., labelled data, inference data) values among the misinformative texts), which strengthens the semantic relatedness between the generated texts and the original texts. Table 1 shows some examples of the common NEs and key terms from the inference data. In summary, the MLM generator substitutes the NEs and key terms, in addition to randomly masked tokens, in the original texts to generate misinformative variants.

Table 1. Top 5 NEs and Key Terms

Named Entities		Key Terms	
Entity	**Frequency**	**Term**	**TF-IDF**
Pfizer	7,108	Vaccine	0.07
Covid	6,021	Covid	0.03
Biden	5,756	Pfizer	0.008
Moderna	2,484	Government	0.008
CDC	1,782	Johnson	0.007

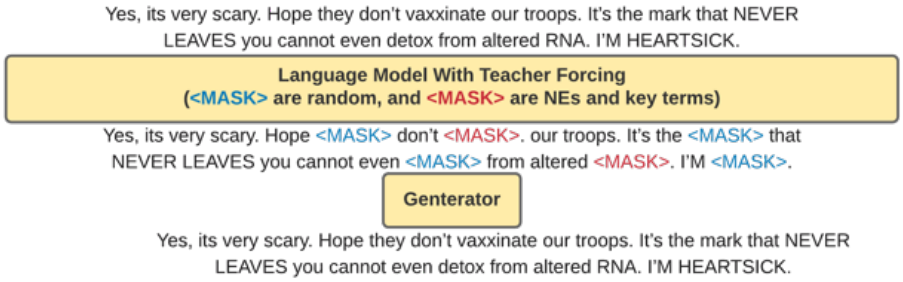

Figure 2. Example of Generated Texts from the MLM Generator

Figure 2 presents the difference between random masking used in generic MLM models and the proposed teacher-forcing mechanism used in our pipeline. It is worth noting that the MLM generator generated exactly the same text as the original message, which demonstrates that the generator is trained to understand the linguistic patterns within the original data. We then perform stratified sampling on the inference data, identifying those which are classified as information, misinformation, and ambiguous data. In this manner, the stratified sampling prevents the downstream discriminator from learning solely based on textual signals relating to the truth behind each text of the sample. A certain number (determined heuristically) of variants are generated for each text in the stratified sample. To further ensure the generated texts are semantically similar to the original text, we develop the discriminator to classify whether a text is *original* or *generated*. The discriminator is a fine-tuned BERT classifier, designed to classify if a piece of text is original (1) or generated (0). This classifier is trained with a sample of the generated texts as discussed above, along with an equal-sized sample of original texts from the inference set. It is worth noting that this set and the texts we select for the MLM generator are mutually exclusive. After the discriminator yields satisfactory classification results, we apply it on the generated texts from the ambiguous data determined by the initial classifier, then select the top-t% generated texts with the highest classification probabilities (t determined heuristically). As discussed in the design of the MLM generator, these generated texts are more misinformative, compared to their original counterparts. These generated texts, termed as *pseudo original* misinformation, are then included in the training set for the enhanced classifier.

3.3. Misinformation Intervention

In the misinformation intervention module, we design and implement another GAN model with GPT-2 as the whole text generator and another fine-tuned BERT model as the whole text discriminator. For the GPT-2 based generator, the training data includes a sample of the inference data, which are classified by the enhanced classifier as true information, as we need to maintain the quality of the generated texts to be less misinformative. To further improve the quality of the generated texts, we also sample a set of true information from the labeled data, which allows for a combination of reliable human and machine labeled texts to act as the foundation for

the generated corrective messages. Similar to the generator in the MLM-GAN model, we select the NEs and key terms and append them as the prefix in each instance in the training set. We design this method to allow the generator to produce texts that are semantically related to the original misinformation (see Section 3.2). The generator is successfully trained once the generated texts it produces are classified as less misinformative in the enhanced classifier. Following this, we treat a sample of classified misinformation from the inference data as input data for the generator. The generator then outputs generated corrective messages based on this input data. The generated corrective messages are then classified using the enhanced classifier to ensure that they are less misinformative (i.e., with lower classification probabilities toward the misinformation class). An example is shown in Table 2.

Table 2. Sample Corrective Message

Original	Generated Message
Gonna kill us all, just you wait and see! #depopulation agenda. … Research before we fume. "Conspiracy Theorist" is looking more and more respectable on a CV nowadays, it's the truth!🤗	Looking forward todays #COVID19 press conference. Grateful for the endless hard work of public health researchers and their teams at CDC. #StayAtHome!

Similar to the discriminator in the MLM-GAN model, we fine-tune a BERT model to identify if the generated corrective message is semantically similar to the true information labeled by humans. We use the same subsample of inference data (labeled as original), which is also used to train the discriminator in the MLM-GAN, combined with a sample of the generated corrective messages (labeled as generated) to fine-tune the discriminator. After the discriminator converges at a satisfactory level, we apply the fine-tuned model to a reserved set of the generated corrective messages (that are not used in the fine-tuning process). We then select the top N (determined heuristically) generated corrective messages based on the classification probabilities of being original, for each misinformation post generated by the whole text generator. These are termed as *candidate corrective messages* for downstream manual evaluation. Two main expected characteristics of the candidate corrective messages are: i) they are less misinformative compared to the original misinformation counterparts, and ii) they are linguistically coherent. These characteristics are evaluated manually to ensure the "human-in-the-loop" nature of the proposed pipeline.

4. Experiment and Results

4.1. Experiment Data and Design

Since COVID-19 is an ongoing event, labeled data regarding misinformation concerning the pandemic is scarce across different social media platforms. Additionally, according to a preliminary analysis of the labeled data, we discover that the extant labeled data are at various quality levels. Thus, we investigated various labeled datasets focusing on COVID-19 misinformation, including COVID-19 Fake News Dataset [26], NewsGuard Coronavirus Misinformation Tracking Dataset [27], COVID Fake News Dataset [28], Poynter CoronaVirusFacts Dataset [29], and CovidLies Dataset [8]. After reviewing these datasets, a sample of 13,947 is included in this study, which contains 6,193 true information posts and 7,754 misinformation posts collected from social media platforms (e.g., Facebook, Twitter) and news articles. Building on findings from prior studies, we retrieve tweets from health care organizations and treat them as true information. Specifically, we programmatically scraped COVID-19 related data from i) articles on World Health Organization (WHO) and Center for Diseases Control and Prevention (CDC) websites (627 true information articles); ii) Tweets from WHO, CDC, and the director of CDC official Twitter accounts (5,580 true information tweets). Conversely, we collect public tweets containing anti-vaccine contents (5,842 misinformation tweets). Each dataset is examined based on qualitative evaluations pertaining to its respective value for either true information or misinformation, including accuracy of statements overall, and accuracy at time of collection versus present day. We then combine all labeled data. In total, we obtain the labeled dataset with a size of 25,996, with 12,400 (47.70%) true information posts and 13,596 (52.30%) misinformation posts. It is worth noting that we perform a 70%/30% split for our training and test datasets.

In order to collect the inference data, we programmatically retrieve Tweets discussing COVID-19, based on Tweet IDs reported from the CoVaxxy project [30]. In order to align the time period covered by the inference data with the labeled dataset, we retrieve the inference data from January 4, 2021 through April 28, 2021. We also perform stratified sampling on days and time of the day to avoid biases toward certain days/hours, producing the inference dataset consisting of 273,000 unlabeled tweets. Within the inference data, we reserve 1,749 (0.6%) tweets used for testing the discriminators in both GAN models, as discussed in Section 3. To implement the design of the pipeline, we set the ratio of generated texts from the discriminator of the MLM-GAN to be 10%, and the number of generated texts from the discriminator of the whole-text GAN to be 2.

4.2. Experiment Results

Since the initial and the enhanced Misinformation Detection Model are both binary classification models, we select the weighted average precision, recall, f1 score and Area Under the Receiver Operator Characteristic Curve (ROC) as the evaluation metrics. The initial model yields high performance, attaining an F1 score of 0.97, and an AUC score of 0.99 against the test set, which outperforms all other classifiers in our experiment. These results are superior compared to the extant detection models toward COVID-19 misinformation (e.g., [1]). However, despite the high level of performance, the initial Misinformation Detection Model shows inadequate performance on the ambiguous data, which are classified incorrectly more often than not, yielding a precision of .97, a recall of .53, an F1 of .68, and AUC score of 0.34, indicating that the initial model correctly classified most of the ambiguous data to be true information (which they actually are), it is less capable of identifying the misinformative posts in them. With the generated examples from the MLM-GAN incorporated into the training of the enhanced Misinformation Detection Model, the AUC score on the ambiguous data improves to 0.86, with the precision, recall, and F1 improved 0.02 – 0.08, respectively (see Table 3). Such improvements suggest that with the more misinformative examples generated by the MLM-GAN, our misinformation detection model is more capable of detecting the borderline misinformation posts.

Table 3. Performances on Ambiguous Data

Model	Precision	Recall	F1	AUC
Initial	.97	.53	.68	.34
Enhanced	.99	.61	.75	.86

The results demonstrate that the detection enhancement in the proposed pipeline significantly improves the robustness of the detection model, with the help of more variance in the training data. Additionally, we quantitatively evaluate the candidate corrective messages against their respective original counterparts, to determine if the candidate corrective messages from the misinformation intervention module are corrective in nature. Since the enhanced Misinformation Detection Model yields reliable results, we leverage it to measure improvement by comparing the classification probabilities of being true information. We randomly sample 317 candidate corrective messages, and the maximum increase in the classification probabilities is 0.995, with an overall average percentage growth of 5,432%. An independent t-test shows that the classification probabilities of the candidate corrective messages and their respective original counterparts are significantly different ($p < 0.05$).

4.3. Manual Evaluation

In order to manually evaluate the efficacy of the candidate corrective messages, we design an evaluation scale (see Figure 4), which measures different levels of required manual intervention toward COVID-19 misinformation. Additionally, this scale provides benchmarking of the misinformation intervention module in measuring how much manual intervention remains toward the *candidate corrective messages*. Specifically, we design the scale levels as follows. Level (i), Exploration/Explanation refers to basic investigations which identifies that misinformation exists in a specific domain and quantitatively analyzes the impact of the dissemination of this false information. The natural progression from this point is Level (ii), Detection, referring to the ability to identify texts as either true information or misinformation in a specific domain. Despite identifying the misinformation, humans must craft responses and/or corrections in a fully manual manner.

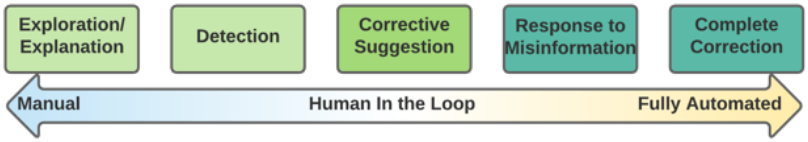

Figure 4. Scale of Required Manual Intervention

Level (iii), Corrective Suggestions, builds on identification of misinformation by providing automatic generation of texts which contain truthful elements pertaining to the misinformative texts. Rather than requiring a fully manual effort to create new responses, human experts can manually modify the generated text to provide the most relevant true information. This step constitutes the first instance of machine-generated texts beginning to lessen the amount of effort required by humans in the effort to combat misinformation. This is progressed as generated texts become Responses to Misinformation (iv). At this level, automatic generation of texts require minimal human intervention to be considered truthful and informative with regards to the initial misinformation. Instances which surpass this require no human intervention and are referred to as Complete Corrections (v). Specifically, generated texts completely correct the initial misinformation and provide the true counterpart information in a fully automated fashion. Most of the extant studies are at the first (exploration/explanation) or the second (detection) level. In order to

provide interventions to combat COVID-19 misinformation, the generated corrective messages need to be at least at the *corrective suggestions* level. We evaluated the *candidate corrective messages* using the aforementioned scale. Due to time constraints, we randomly sample the top 0.5% (500 data points) from the inference data, by their classification probabilities from the enhanced classifier. We also manually review this sample to ensure they are indeed COVID-19 misinformation contents in order to evaluate the intervention module in isolation. We then leverage the proposed misinformation intervention module (see Section 3.3) to generate a maximum of 15 candidate corrective messages for each misinformation post in the evaluation sample. Three raters from different backgrounds independently evaluate each text, against the levels in the proposed scale (see Figure 4).

. **Table 4. Example of Candidate Corrective Messages in Different Categories**

Original Misinformation Text	Generated Corrective Messages
Example of Complete Correction	
Finally people are waking up from this slavery!!! People are a lot more powerful then these global fascists and their plandemic!! #stopmasks #DemocratsAreDestroyingAmerica #FightDictators #NOVaccinePassports #NoVaccine	#WearAMask over your nose and mouth to stop the spread of this virus. #CovidVaccine #WearAMask
Example of Response to Misinformation	
Ooops! He let it slip 🙈 New Variants Being "DEVELOPED" Plandemic/Scamdemic at it's best.	Here are 5 tips that help you stay safe to protect yourself and others from #COVID19: clean hands wear a face mask stay distant cough, sneeze safely
Example of Corrective Messages	
Valid Claims Natural Remedies Prevent, Treat or Cure Covid-19 Are Now Being Censored by the FDA As folks realize that COVID is a #Plandemic that is being used to terrorize us into slavery, the elites and social media oligarchs are in an absolute panic.	The U.S. is using simulation exercises to prepare for and prevent real-world #COVID19.

The raters conducted a training and discussion phase on 50 randomly sampled *candidate corrective messages* to ensure a consensus of the scale levels. The sampled evaluation data are rated as either: corrective suggestion, responses to misinformation, or complete correction, as described above. All posts that cannot be classified as any of the above three levels are categorized as detection, according to the reliable performance of the enhanced Misinformation Detection Model. Our initial results show that approximately 75% of cases had at least one rater believing the candidate corrective messages were at least Corrective Suggestions, and our raters were in complete agreement in approximately 18% of the cases that the *candidate corrective messages* from the evaluation data were Corrective Suggestions or higher. The overall agreement among all raters is 43%. We discuss the possible reasons for the moderate agreement level in Section 5. On average our raters view approximately 52% of the *candidate corrective messages* as being at least corrective suggestions, which indicates that more than half of the *candidate corrective messages* are of a corrective nature. Additionally, we calculate intra-rater reliability in a pairwise manner between each rater using Cohen's Kappa statistic, which ranges from 0, signaling rating agreement as purely random, to 1 which implies perfect agreement. Our Fleiss Kappa statistic is 0.45, which signifies moderate agreement among raters (0.41 – 0.80). Combined with the statistics reported above, this value suggests that the proposed analytical pipeline incorporates both human and machine intelligence in the intervention loop. Also, the low quality of the extant labeled data on COVID-19 misinformation suggests that labeling/rating is a difficult task, thus we believe the results from this study are of sufficient reliability. Table 4 presents examples of the corrective messages passed the human ratings, and the respective original counterparts.

5. Lessons Learned

Both the automated and manual evaluation results provide clear evidence that the proposed analytical pipeline is capable of generating corrective messages, which not only exhibits superior detection effectiveness, but also extends current studies concerning misinformation on social media to a new level. In addition, we learned several lessons from the experiment and its results reported in Section 4.

The manual evaluation results are conclusive in displaying state-of-the-art advancements of human-in-the-loop misinformation intervention, albeit with

moderate intra-rater agreement. The reasoning behind this phenomenon is three-fold. First, the COVID-19 pandemic is ongoing and changing rapidly. Therefore, prior and stable information, which is required in a qualitative review, is nearly non-existent [1]. Second, the expertise of human raters are disjointed, consisting of data science, data engineering, and the medical profession respectively. Levels of disagreement among raters evaluating the corrective suggestions stem from the differing subject matter perspectives among raters. Third, prior studies (e.g., [1], [11]) have suggested that human knowledge is insufficient in detecting COVID-19 misinformation, which calls for (semi-)automatic decision tools. It is also highlighted in previous studies that during an ongoing public health event (e.g., COVID-19), it is difficult to identify the experts and to seek expert opinions [3]. We believe that moderate agreement despite these differences in backgrounds serves as an unambiguous sign of success with our process. Furthermore, the proposed analytical pipeline enables more timely responses to misinformation compared to manual responses, which can reduce the effect of misinformation despite the coherence level of the corrective messages [31].

Our proposed analytical pipeline was initially theorized as a means to aid human-in-the-loop efforts to intervene against COVID-19 misinformation on social media. However, after reviewing the manual evaluation results, we discovered that the proposed pipeline can be used as a training tool for the individuals and related organizations who combat COVID-19 misinformation. Our pipeline is capable of collecting large quantities of both misinformative and informative texts to help human users understand the differences between the two. Additionally, the misinformation intervention module provides examples of how to alter or update misinformation in order to correct these texts, which largely improves the performance of non-domain experts. By developing an intelligent system that serves as a model for their imminent responsibilities, we can help to fully train the workforce to combat misinformation, which is not limited to the context of COVID-19, without requiring as much domain knowledge.

6. Conclusion

COVID-19 misinformation on social media has become a severe issue in the ongoing global pandemic, jeopardizing public health by hindering social media users from perceiving valuable information regarding treatments and best practices. Extant studies have focused on the impact and motivation of COVID-19 misinformation, and largely ignored intervening against them. In this study, we propose an analytical pipeline, which not only yields superior misinformation detection results via advanced detection models and GAN based data augmentation tools, but also provides corrective suggestions to assist human users in combating COVID-19 misinformation. The intensive automatic and manual evaluation results showed superior efficacy of the proposed pipeline toward COVID-19 misinformation detection and intervention. Additionally, this pipeline is generalizable to misinformation in other contexts if the models are fine-tuned with relevant texts. Moreover, the proposed pipeline can be used to bridge the gap in the disjoint expertise in human analysts, and to train non-domain experts concerning misinformation intervention.

We acknowledge that this study comes with several limitations, which may point to possible directions for future research. Firstly, the proposed pipeline is a decision support system, which is not able to generate complete corrective information to combat COVID-19 misinformation. Even with the help of the latest developments in transformer-based models, such as GPT-3, we still believe that the pipeline is most suitable as an intelligent assistant/educational tool. Secondly, although the transformer-based models (e.g., BERT, GPT-2) are capable of capturing the linguistic nuances in social media texts, we believe additional measures, such as semantic similarity between the generated texts and their original counterparts should be considered, in order to improve the relatedness between them. Thirdly, we believe that improved quality of the labeled training data can lead to improved semantic correctness and scientific precision of the generated corrective messages. Thus, we plan to conduct intensive manual labeling improve the label accuracy. Fourthly, we believe that the performance gap between the initial and enhanced detection models can be partially attributed to the writing style differences between true and false information. More data at various coherence levels can help with the generalizability of the proposed pipeline. Lastly, the misinformation can be further categorized according to different topics (e.g., vaccine, masks), and different pipelines can be developed to fight misinformation under each topic. Another way to address this limitation is possibly incorporate the reinforcement learning mechanism in the design of the proposed analytical pipeline [32].

7. References

[1] J. Choudrie, S. Banerjee, K. Kotecha, R. Walambe, H. Karende, and J. Ameta, "Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study," Computers in Human Behavior, vol. 119, no. January, p. 106716, 2021.
[2] A. K. M. N. Islam, S. Laato, S. Talukder, and E.

Sutinen, “Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective,” Technological Forecasting and Social Change, no. 159, 2020.
[3] E. K. Vraga and L. Bode, “Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation,” Political Communication, vol. 37, no. 1, pp. 136–144, 2020.
[4] A. Mitchell, J. B. Oliphant, and E. Shearer, “About Seven-in-Ten U.S. Adults Say They Need to Take Breaks From COVID-19 News,” Pew Research Center Research Report, 2020. [Online]. Available: https://www.journalism.org/2020/04/29/about-seven-in-ten-u-s-adults-say-they-need-to-take-breaks-from-covid-19-news/.
[5] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, “Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA,” Nature Human Behaviour, vol. 5, no. 3, pp. 337–348, 2021.
[6] L. Singh et al., “A first look at COVID-19 information and misinformation sharing on Twitter,” arXiv, 2020.
[7] R. Kouzy et al., “Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter,” Cureus, vol. 12, no. 3, 2020.
[8] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, “COVIDLies: Detecting COVID-19 Misinformation on Social Media,” in Proceedings of the 1st Workshop on NL for COVID-19 (Part 2) at EMNLP 2020, 2020.
[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” in The International Conference on Learning Representations 2020, 2019, pp. 1–43.
[10] S. Tasnim, M. Hossain, and H. Mazumder, “Impact of rumors and misinformation on COVID-19 in Social Media,” Journal of Preventive Medicine and Public Health, vol. 53, no. 3, pp. 171–174, 2020.
[11] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, “An exploratory study of COVID-19 misinformation on Twitter,” Online Social Networks and Media, vol. 22, no. January, 2021.
[12] N. M. Krause, I. Freiling, B. Beets, and D. Brossard, “Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19,” Journal of Risk Research, vol. 23, no. 7–8, pp. 1052–1059, 2020.
[13] I. Goodfellow et al., “Generative adversarial networks,” 2014.
[14] Z. Barua, S. Barua, S. Aktar, N. Kabir, and M. Li, “Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation,” Progress in Disaster Science, vol. 8, p. 100119, 2020.
[15] R. E. Irwin, “Misinformation and de-contextualization: International media reporting on Sweden and COVID-19,” Globalization and Health, vol. 16, no. 1, pp. 1–12, 2020.
[16] S. Evanega, M. Lynas, J. Adams, and K. Smolenyak, “Coronavirus misinformation: quantifying sources and themes in the COVID-19 ‘infodemic,’” JMIR Preprints, pp. 1–13, 2020.
[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” 2013.
[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, no. Mlm, pp. 4171–4186, 2019.
[19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” pp. 1–18, 2019.
[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” OpenAI Blog, vol. 1, no. May, pp. 1–7, 2020.
[21] P. Niewinski, M. Pszona, and M. Janicka, “GEM: Generative Enhanced Model for adversarial attacks,” in Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), 2019, pp. 20–26.
[22] X. Wu, T. Zhang, L. Zang, J. Han, and S. Hu, “Mask and Infill : Applying Masked Language Model to Sentiment Transfer,” 2019.
[23] Y. Zhang, Z. Gan, and L. Carin, “Generating text via adversarial training,” in NIPS workshop on Adversarial Training, 2016, vol. 21, pp. 1–6.
[24] A. A. Irissappane, H. Yu, Y. Shen, A. Agrawal, and G. Stanton, “Leveraging GPT-2 for Classifying Spam Reviews with Limited Labeled Data via Adversarial Training,” pp. 1–26, 2020.
[25] A. Lamb, A. Goyal, Y. Zhang, S. Zhang, A. Courville, and Y. Bengio, “Professor forcing: A new algorithm for training recurrent networks,” in 29th Conference on Neural Information Processing Systems (NIPS 2016), 2016, pp. 4608–4616.
[26] P. Patwa et al., “Fighting an Infodemic: COVID-19 Fake News Dataset.” 2020.
[27] NewsGuard, “Coronavirus Misinformation Tracking Dataset,” 2021. [Online]. Available: https://www.newsguardtech.com/coronavirus-misinformation-tracking-center/. [Accessed: 22-Apr-2021].
[28] S. Banik, “COVID Fake News Data.” 2020.
[29] Poynter, “CoronaVirus Facts Dataset,” 2021. [Online]. Available: https://www.poynter.org/coronavirusfactsalliance/. [Accessed: 04-Mar-2021].
[30] M. R. DeVerna et al., “CoVaxxy: A Collection of English-language Twitter Posts About COVID-19 Vaccines,” 2021.
[31] N. Walter and R. Tukachinsky, “A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?,” Communication Research, vol. 47, no. 2, pp. 155–177, 2020.
[32] J. Xu, X. Ren, J. Lin, and X. Sun, “Diversity-promoting GaN: A cross-entropy based generative adversarial network for diversified text generation,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pp. 3940–3949, 2020.